Sentiment Analysis: A Survey

Sneha S Mulatkar

***Abstract*—Traditional approaches to sentiment classification rely on lexical features, syntax-based features or a combination of the two. Word senses used as features show promise, we also examine the possibility of using similarity metrics defined on WordNet to address the problem of not finding a sense in the training corpus. Different methods of sentiment classification are also described in it. Thus is provides a broad way of analyzing the methods and making conclusion provided by these methods.**

Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, by analyzing a large numbers of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or writer with respect to some topic of a document. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today.

Index Terms— **Corpus, Gloss, Synset ,WordNet.**

I. INTRODUCTION

Sentiment Analysis (SA) is the task of prediction of opinion in text. Sentiment classification deals with tagging text as positive, negative or neutral from the perspective of the speaker/writer with respect to a topic. Classification task for output labels as positive and negative [5]. Traditional supervised approaches for SA have explored lexeme and syntax-level units as features.

Approaches using lexeme-based features use bag-of- Words or identify the roles of different parts-of-speech (POS) like adjectives.

WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents. For example, the words night, nighttime, and dark constitute a single synset that has the following gloss: the time after sunset and before sunrise while it is dark outside. Synsets are connected to one another through explicit semantic relations. Some of these relations (hypernym, hyponym for nouns, and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies. For example, tree is a kind of plant, tree is a hyponym of plant, and plant is a hypernym of tree.

It explores incorporation of semantics in a supervised sentiment classifier. We use the synsets in WordNet as the feature space to represent word senses [5].

Manuscript received March 21, 2014.
Sneha Mulatkar, Information Technology, Mumbai University, Navi Mumbai, India, 9870751523, (e-mail: sneham.29@gamil.com).

II. CHALLENGES FOR SENTIMENT ANALYSIS

Sentiment Analysis approaches aim to extract positive and negative sentiment bearing words from a text and classify the text as positive, negative or else objective if it cannot find any sentiment bearing words. [2]In this respect, it can be thought of as a text categorization task. In text classification there are many classes corresponding to different topics whereas in Sentiment Analysis we have only 3 broad classes. Thus it seems Sentiment Analysis is easier than text classification which is not quite the case. The general challenges can be summarized as:

A. Implicit Sentiment and Sarcasm

A sentence may have an implicit sentiment even without the presence of any sentiment bearing words. Consider the following examples.

How can anyone sit through this movie?

One should question the stability of mind of the writer who wrote this book.

Both the above sentences do not explicitly carry any negative sentiment bearing words although both are negative sentences. Thus identifying semantics is more important in SA than syntax detection.

B. Domain Dependency

There are many words whose polarity changes from domain to domain. Consider the following examples.

The story was unpredictable.

The steering of the car is unpredictable.

Go read the book.

In the first example, the sentiment conveyed is positive whereas the sentiment conveyed in the second is negative. The third example has a positive sentiment in the book domain but a negative sentiment in the movie domain.

C. Thwarted Expectations

Sometimes the author deliberately sets up context only to refute it at the end. Consider the following example:

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance.

However, it can't hold up. Inspite of the presence of words that are positive in orientation the overall sentiment is

negative because of the crucial last sentence.

D. Pragmatics

It is important to detect the pragmatics of user opinion which may change the sentiment thoroughly. Consider the following examples:
I just finished watching Barca DESTROY Ac Milan
That final completely destroyed me.
Capitalization can be used with subtlety to denote sentiment. The first example denotes a positive sentiment whereas the second denotes a negative sentiment.

E. Subjectivity Detection

This is to differentiate between opinionated and non-opinionated text. This is used to enhance the performance of the system by including a subjectivity detection module to filter out objective facts. But this is often difficult to do.

Consider the following examples:

I hate love stories.

I do not like the movie "I hate stories".

The first example presents an objective fact whereas the second example depicts the opinion about a particular movie.

F. Negation

Handling negation is a challenging task in SA. Negation can be expressed in subtle ways even without the explicit use of any negative word. A method often followed in handling negation explicitly in sentences like "I do not like the movie", is to reverse the polarity of all the words appearing after the negation operator (like not). But this does not work for "I do not like the acting but I like the direction". So we need to consider the scope of negation as well, which extends only till but here. So the thing that can be done is to change polarity of all words appearing after a negation word till another negation word appears. But still there can be problems.

For example, in the sentence "Not only did I like the acting, but also the direction", the polarity is not reversed after "not" due to the presence of "only". So this type of combinations of "not" with other words like "only" has to be kept in mind while designing the algorithm.

III. TRADITIONAL METHOD VS MODERN METHOD

Web content mining is intended to help people discover valuable information from large amount of unstructured data on the web. Movie review mining classifies movie reviews into two polarities: positive and negative. [2] As a type of sentiment-based classification, movie review mining is different from other topic-based classifications. The main objective of this work is to classify a large number of opinions using web-mining techniques into bipolar orientation (i.e. either positive or negative opinion). Such kind of classification could help consumers in making their purchasing decisions.

Research results along this line can lead to users' reducing the time on reading threads of text and focusing more on analyzing summarized information. Review mining can be potentially applied in constructing information presentation.

FEATURES FOR SENTIMENT ANALYSIS

Feature engineering is an extremely basic and essential task for Sentiment Analysis. Converting a piece of text to a feature vector is the basic step in any data driven approach to SA. In the following section we will see some commonly used features used in Sentiment Analysis and their critiques.

TERM PRESENCE VS. TERM FREQUENCY

Term frequency has always been considered essential in traditional Information Retrieval and Text Classification tasks. But Pang-Lee et al. (2002) found that term presence is more important to Sentiment analysis than term frequency. [5] That is, binary-valued feature vectors in which the entries merely indicate whether a term occurs (value 1) or not (value 0). This is not counter intuitive as in the numerous examples we saw before that the presence of even a single string sentiment bearing words can reverse the polarity of the entire sentence. It has also been seen that the occurrence of rare words contain more information than frequently occurring words, a phenomenon called Hapax Legomena.

TERM POSITION

Words appearing in certain positions in the text carry more sentiment or weightage than words appearing elsewhere. This is similar to IR where words appearing in topic Titles, Subtitles or Abstracts etc are given more weightage than those appearing in the body. Although the text contains positive words throughout, the presence of a negative sentiment at the end sentence plays the deciding role in determining the sentiment. Thus generally words appearing in the 1st few sentences and last few sentences in a text are given more weightage than those appearing elsewhere.

IV. DIFFERENT METHODS

Corpus

We need a collection of movie reviews that include both positive reviews and negative reviews. Good corpus resources should have good review quality, available metadata, easy spidering, and reasonably large number of reviews and products. There are 1,400 text files in total with 700 labeled as positive reviews and the rest 700 labeled as negative reviews. They were originally collected from IMDB (Internet Movie Database)[4] archive of review newsgroups at http://reviews.imdb.com/Reviews. The ratings were removed. The rating decision was made in order to transform a 4-star or 5-star rating-system reviews into positive and negative reviews. Moreover, the data were examined manually to ensure quality. A few non- English and incomplete reviews were removed [4]. Misclassified reviews based on sentimental judgment were also corrected.

A. N-gram Classifiers

In the light that n-gram models provide one of the best performance in text classification in general, we selected n-gram models as supervised approach, which represent text documents by word tuples. [4] The tool is implemented with classification algorithms based on n-gram (unigram, bigrams, and tri-grams) features. Several options are given to adapt classification models, such as adding stop-word lists. The stop-word lists can be in unigram, bi-grams, and tri-grams forms. It is risky to over-filter data, for we may remove important information along with unwanted data. Therefore, we captured a large number of features at the beginning, and then tried

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts. True Type 1 or Open Type fonts are required. Please embed all fonts, in particular symbol fonts, as well, for math, etc. multiple sets of features to select the one that best suits our problems.

B. Semantic Orientation (SO) approach

Based on parts-of-speech in the parsed output, two-word phrases were then selectively extracted. Only two-word phrases conforming to certain patterns were extracted for further processing. Adjective or adverb in the patterns provides subjectivity, while the other word provides context [5]. The following table summarizes five patterns used in the extraction of phrases.

Two-word phrase patterns

First word	Second word
Adjective	Noun
Adverb	Adjective
Adjective	Adjective
Noun	Adjective
Adverb	Verb

The next step was to determine the semantic orientation of a phrase' SO (phrase) according to Formula (1) [5] hits (·) denotes the number of pages returned for a query consisting of phrase · from a search engine.

For example,

Hits ('poor') represents the number of pages returned for a search query 'poor'. When there are both phrase and 'excellent' (or 'poor') connected by NEAR operator in the parameter of hits function, it defines the similarity between phrase and 'excellent' (or 'poor') [5]. In other words, the similarities were measured with co occurrences of the phrases and 'excellent' (or 'poor').

SO (phrase) =

$$\text{Log}_2 = \left\{ \frac{\text{hits(phrase NEAR “excellent”) hits(“poor”)}}{\text{hits(phrase NEAR “poor”) hits(“excellent”)}} \right\}$$

A phrase's semantic orientation would be positive if it is associated more strongly with "excellent" than "poor" and would be negative if it is associated more strongly with "poor" than "excellent". Finally, a review's semantic orientation was calculated by averaging the SO values of all the extracted phrases in it. The movie is recommended to watch if its average semantic orientation exceeds a threshold and is not recommended if otherwise.

C. Rating decisions

A movie was rated as one of the five categories [4]. If we are to group them into recommended and not recommended, we should set a dividing line to separate the data. It is obvious that the data are positively skewed. Therefore, it is a subjective decision as to whether a particular rating should fall under positive or negative categories, especially for ratings with neutral tone (such as B and C). Some ratings are in different forms and some are even missing, making them incomparable across different rating systems of different reviewers [4]. With a five-star rating system, reviews given four stars and up were considered positive while reviews given two stars and below were negative. With four-star rating system, reviews given three stars and up were positive while reviews given one star and below were considered negative.

Reviews that fall in neutral range are discarded. Therefore, we decided to ignore the neutral ratings, and group ratings into two main categories as follows:

The rating system of movie justice was comparable to a five-star rating system (A, B, C, D, and F comparing to 5- , 4-, 3-, 2-, and 1-star) [4] Accordingly, movies rated in A and B ranges received positive reviews and those rated in D and F ranges received negative reviews, but movie reviews in C range were discarded.

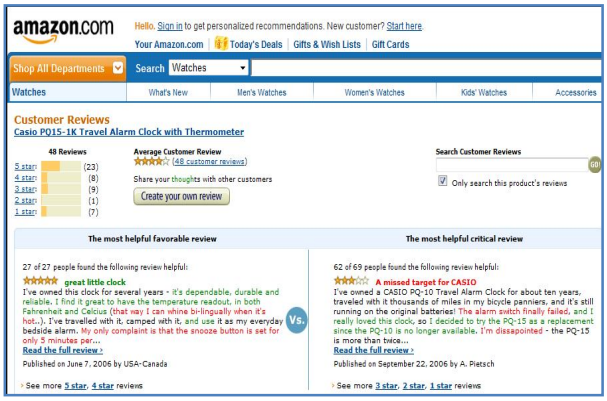

Fig 1: Demo-Instant page analysis

Given a URL, it automatically identifies opinions on the page. Green: +ve, and red: -ve

D. Polysemy Reduction Approaches

Polysemy has been addressed in two main fields [1]: in Information Retrieval (IR), to increase effectiveness of IR systems and in word sense disambiguation (WSD), where the focus is on complementary polysemy and on how to identify the meaning of polysemous words in a given context. IR approaches aim to produce more coarse-grained lexical resources of existing fine-grained ones such as WordNet, i.e. polysemy reduction. WSD approaches focus on the recognition and identification of the intended meaning of ambiguous polysemous words using the surrounding context.

In polysemy reduction, the senses are clustered such that each group contains related polysemous words .They are

called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged.

Polysemy reduction approaches typically rely on the application of some detection rules such as [1]:

If S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset.

However, applying this rule may wrongly result in merging two different senses as in the following example:

Smoke, smoking -- a hot vapor containing fine particles of carbon being produced by combustion; "the fire produced a tower of black smoke that could be seen for miles"

Smoke, smoking -- the act of smoking tobacco or other substances; "he went outside for a smoke"; "smoking stinks"

Polysemy Reduction Rules:

Let S1 and S2 be two synsets in WordNet, then S1 and S2 can be merged if they fulfill at least one of the following rules:

Rule 1: If S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words.

Rule 2: If S1 and S2 are two synsets with the same hypernym or one of them is a direct hypernym of the other.

Rule 3: if S1 and S2 have the same direct hyponym synset or one is a direct hyponym of the other.

Rule 4: If S1 and S2 have the same coordinate terms (i.e., there exist a synset S3 such that S1 and S3 share a direct hypernym, and S2 and S3 share a direct hypernym).

Rule 5: If S1 and S2 have the same antonym.

Rule 6: S1 and S2 a have the same pertainym.

Rule 7: If S1 and S2 have similar to terms in common (i.e., there exist a synset S3 such that S1 is similar to S3, and S2 is similar to S3)

Rule 8: If S1 and S2 have related to terms in common (i.e., there exist a synset S3 such that S1 is related to S3, and S2 is related to S3).

E. Synset Replacement Algorithm

Using WordNet senses provides an opportunity to use similarity-based metrics for WordNet to reduce the effect of unknown features. If a synset encountered in a test document is not found in the training corpus, it is replaced by one of the synsets present in the training corpus [5]. The substitute synset is determined on the basis of its similarity with the synset in the test document. The synset that is replaced is referred to as an unseen synset as it is not known to the trained model.

For example, consider excerpts of two reviews, the first of which occurs in the training corpus while the second occurs in the test corpus.

1. “In the night, it is a lovely city and...”

2. “The city has many beautiful hot spots for tourist.

The synset of ‘beautiful’ is not present in the training corpus. We evaluate a similarity metric for all synsets in the training corpus with respect to the sense of beautiful and find that the sense of lovely is closest to it.

Hence, the sense of beautiful in the test document is replaced by the sense of lovely which is present in the training corpus.

V. APPLICATIONS OF SENTIMENT ANALYSIS

Word of mouth (WOM) is the process of conveying information from person to person and plays a major role in customer buying decisions. In commercial situations [2], WOM involves consumers sharing attitudes, opinions, or reactions about businesses, products, or services with other people. WOM communication functions based on social networking and trust. People rely on families, friends, and others in their social network. [2] Research also indicates that people appear to trust seemingly disinterested opinions from people outside their immediate social network, such as online reviews. This is where Sentiment Analysis comes into play. Growing availability of opinion rich resources like online review sites, blogs, social networking sites have made this “decision-making process” easier for us. With explosion of Web 2.0 platforms consumers have a soapbox of unprecedented reach and power by which they can share opinions. Major companies have realized these consumer voices affect shaping voices of other consumers.

Sentiment Analysis thus finds its use in Consumer Market for Product reviews, marketing for knowing consumer attitudes and trends, Social Media for finding general opinion about recent hot topics in town, Movie to find whether a recently released movie is a hit. Pang-Lee et al. (2002) broadly classifies the applications into the following categories.

a) Applications to Review-Related Websites

Movie Reviews, Product Reviews etc.

b)Applications as a Sub-Component Technology

Detecting antagonistic, heated language in mails, spam detection, context sensitive information detection etc.

c) Applications in Business and Government Intelligence

Knowing Consumer attitudes and trends

d) Applications across Different Domains

Knowing public opinions for political leaders or their notions about rules and regulations in place.

VI. ANALYSIS

Sentiment analysis - a discipline of information retrieval – the opinion mining (OM). OM analyzes the characteristics of opinions, feelings and emotions that are expressed in textual or spoken data with respect to a certain subject. Subtask of sentiment analysis - categorization on the basis of certain polarities - the sentiment polarity identification.

Sentiment Analysis (SA) is the task of prediction of opinion in text. Sentiment classification deals with tagging text as positive, negative or neutral from the perspective of the speaker/writer with respect to a topic. Classification task for output labels as positive and negative.

Term frequency has always been considered essential in traditional Information Retrieval and Text Classification tasks. It is found that term presence is more important to Sentiment analysis than term frequency. It has also been seen that the occurrence of rare words contain more information than frequently occurring words.

To overcome this problem, information contained in term glosses – explanatory text accompanying each term – can be explored to infer term orientation, based on the assumption that a given term and the terms contained in its gloss are likely to indicate the same polarity

The method of lexicon expansion is proposed where terms are assigned positive or negative opinions based on the existence of terms known to carry opinion content found on the term gloss.

The idea is also seen by using supervised learning methods for extending a lexicon by exploring gloss information, yielding positive accuracy improvements used in SentiWordNet opinion.

N-gram Classifiers: The tool is implemented with classification algorithms based on n-gram (unigram, bigrams, and tri-grams) features. Stop word are removed. But there occurs the risk of over filtering and removing the important features.

Semantic Orientation (SO) approach: A phrase's semantic orientation would be positive if it is associated more strongly with "excellent" than "poor" and would be negative if it is associated more strongly with "poor" than "excellent". Finally, a review's semantic orientation was calculated by averaging the SO values of all the extracted phrases in it. This involves mathematical calculations.

Polysemy reduction approach: In polysemy reduction, the senses are clustered such that each group contains related polysemous words .They are called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged.

Table 1: Comparison Table

Semantic Orientation(SO) approach	**Polysemy Reduction Approaches**	**Synset Replacement Algorithm**
Two word phrases were selectively extracted. This is done for further processing.	This approach have some rules defined which is used to classify that the text is positive or negative	This approach is mainly used to replace those word in the text which are present in the training corpus
The combination of words are extracted like Verb-adjective Adjective-adjective Adverb-verb Adjective-noun	In polysemy reduction, the senses are clustered such that each group contains related polysemous words .They are called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged.	In this approach a word is replace with its synset.
The next step is to determine the SO value by formula $\log_2$=(hits(phrase Near"excellent"hits("poor")/hits)(phrases Near "poor")hits(excellent)) A phrase's semantic orientation would be positive if it is associated more strongly with "excellent" than "poor" and would be negative if it is associated more strongly with "poor" than "excellent". Finally, a review's semantic orientation was calculated by averaging the SO values of all the extracted phrases in it.	The polysemous can be merged by the given rule: 1)S1 and S2 contain the same words. 2) If S1 and S2 are two synsets with the same hypernym. 3) If S1 and S2 have the same direct hyponym synset 4) If S1 and S2 have the same coordinate terms. 5) If S1 and S2 have the same antonym. 6)S1 and S2 a have the same pertainym. 7)If S1 and S2 have similar to terms in common. 8)If S1 and S2 have related to terms in common	In synset replacement algorithm The word which is not present in the training corpus is replaced with the word which is present in the training corpus. Only that word is selected which have same synset also the meaning of the sentence should not change
This process involves a mathematical calculation of the ratings. Suppose we are using this system for movie recommendation .the movie ABC is receive a positive response, there may be the possibility that the adults like the movie but due to violence the children are not recommended to watch it.So there the system may fail	This approach by using the predefined rule for reduction is better than the other described here. only care to be taken is during the formation of the homograph clusters.	In this method there are chances that the words may be same ,with same synset but used in different sense. There is limitation to the training corpus. those limitedwords cannot fulfill requirements The system may fail.

VII. CONCLUSION

This study suggested that human language is very subtle and some meanings conveyed were not captured by the existing patterns. Good movies containing violent or unhappy scenes were often recognized incorrectly. Not only does it deal with classification of personal opinions, but diverse opinions from product reviews as well. Due to the sparsity of words in reviews, it is difficult for supervised learning approach to use bag-of-words features. Review mining is a very challenging issue for semantic orientation techniques. The findings of this study not only advance the research on movie review mining but also in opinion mining.

ACKNOWLEDGMENT

I owe a great many thanks to a many people who helped and supported me. My deepest thanks to the Guide of the project for guiding and correcting various documents of mine with attention and care..I express my thanks to the Principal of, Pillai Institute of Information Technology, New Panvel for extending his support. My deep sense of gratitude to IJIRCST for giving me an opportunity for publishing my manuscript.

REFERENCES

[1] "A systematic Approach towards the Solution of the Polysemy Problem in Natural Language Processing" ,Abed Alhakim Freihat April 2011.

[2] "Approach to Sentiment Analysis: Analytical Categories and Issues of Automation", Repindex.

[3] "HarnessingWordNet Senses for Supervised Sentiment Classification",Balamurali A,Aditya Joshi,Pushpak Bhattacharyya IITB-Monash Research Academy, IIT Bombay Dept. of Computer Science and Engineering, IIT Bombay.

[4] "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches",Pimwadee Chaovalit Department of Information Systems University of Maryland, Baltimore County Lina Zhou Department of Information Systems University of Maryland, Baltimore County.

[5] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., 2002.

[6] "Sentiment Analysis, Indian Institute of Technology",Subhabrata Mukherjee, Bombay Department of Computer Science and Engineering, June 29, 2012.

[7] "Sentiment Classification in Movie Reviews",An Approach Using Subjectivity Filtering Daniel Pomerantz, McGill University.

[8] "Sentiment Classification of Reviews Using SentiWordNet", Ohana, B., Tierney, B.: Sentiment classification of reviews using SentiWordNet. 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland.

Sneha Mulatkar –I have completed my Bachelors in Information Technology and pursuing my master in Information Technology.I am doing my research work for my masters in natural language processing under the guidance of my project guide.